Modeling relationships in referential expressions with compositional modular networks

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K.

Presenter: Minzhi Liao
Hooman Hashemi
Introduction
  Motivation
  Related work
Model in the paper
  Expression parsing
  Localization module
  Relationship module
Experiments
  Synthetic dataset
  Visual Genome dataset
  Google-Ref dataset
  Visual-7W dataset
Outline

Introduction
  Motivation
  Related work
Model in the paper
  Expression parsing
  Localization module
  Relationship module
Experiments
  Synthetic dataset
  Visual Genome dataset
  Google-Ref dataset
  Visual-7W dataset
Motivation

Problem: given a referential expression and image localize the entities
Main idea: dynamic assembling network architecture with modules for simple tasks
Main idea: dynamic assembling network architecture with modules for simple tasks
Classification

classify : \( \text{Image} \times \text{Attention} \rightarrow \text{Label} \)

Re-attention

re-attend : \( \text{Attention} \rightarrow \text{Attention} \)
Main idea: dynamic assembling network architecture with modules for simple tasks
Localizing entities based on arbitrary natural language expression is a challenging problem. Previous work either treat referential expression holistically, or relies on a fixed set of entity and relationship categories.
Model in the paper

(a) language representation

- word sequence \( \{w_i\} \)
  - the man riding a black bicycle
  - word embedding sequence \( \{e_i\} \)
  - 2-layer bidirectional LSTM
  - concatenated state \( [h_f] \)
  - fully connected layer 1
  - fully connected layer 2
  - fully connected layer 3
  - subject relationship
  - relationship
  - object
  - three attention weights over each word \( \{a_{i,subj}\}, \{a_{i,obj}\}, \{a_{i,rel}\} \)
  - subject
  - relation
  - object
  - weighted average
  - \( q_{subj} \)
  - \( q_{rel} \)
  - \( q_{obj} \)

(b) localization module

- image region \( b \)
  - language representation of subject or object \( q_{bc} \)
  - CNN
  - local visual feature
  - local spatial feature
  - concatenate
  - fully connected layer
  - element-wise multiplication
  - 12-normalization
  - output unary score \( s_{rel} \)

(c) relationship module

- image region \( b_1 \)
  - b1 local spatial feature
  - concatenate
  - fully connected layer
  - element-wise multiplication
  - 12-normalization
  - output pairwise score \( s_{rel} \)
Expression parsing

the apple on top of the bookshelf
the apple on top of the bookshelf

Results from syntactic parsers not always correspond to intuitive visual representations
Expression parsing

(a) language representation

- word sequence \( \{w_t\} \)
- the man riding a black bicycle

- word embedding sequence \( \{e_t\} \)

- 2-layer bidirectional LSTM

- concatenated state \( \{h_t\} \)

- fully connected layer 1 softmax
- fully connected layer 2 softmax
- fully connected layer 3 softmax

- three attention weights over each word
  \( \{a_{t,subj}\}, \{a_{t,rel}\}, \{a_{t,obj}\} \)

- subject
- relationship
- object

- weighted average

- \( q_{subj} \)
- \( q_{rel} \)
- \( q_{obj} \)

- embedding

- \( e_t \)

- Bi-LSTM

- \( h_t^{(1,fw)} \) \
- \( h_t^{(1,bw)} \) \
- \( h_t^{(2,fw)} \) \
- \( h_t^{(2,bw)} \)

- \( h_t = [h_t^{(1,fw)}, h_t^{(1,bw)}, h_t^{(2,fw)}, h_t^{(2,bw)}] \)
Expression parsing

The man riding a black bicycle

word sequence \( \{w_t\} \)

word embedding sequence \( \{e_t\} \)

2-layer bidirectional LSTM

concatenated state \( \{h_t\} \)

three attention weights over each word \( \{a_{t,subj}\}, \{a_{t,rel}\}, \{a_{t,obj}\} \)

subject \( a_{t,subj} \)

relationship \( a_{t,rel} \)

object \( a_{t,obj} \)

weighted average

\( a_{t,subj} = \exp (\beta_{subj}^T h_t) / \sum_{\tau=1}^{T} \exp (\beta_{subj}^T h_\tau) \)

\( a_{t,rel} = \exp (\beta_{rel}^T h_t) / \sum_{\tau=1}^{T} \exp (\beta_{rel}^T h_\tau) \)

\( a_{t,obj} = \exp (\beta_{obj}^T h_t) / \sum_{\tau=1}^{T} \exp (\beta_{obj}^T h_\tau) \)
Expression parsing

(a) Language representation

- word sequence \( \{w_t\} \)
  - the man riding a black bicycle

- word embedding sequence \( \{e_t\} \)

- 2-layer bidirectional LSTM

- concatenated state \( \{h_t\} \)

- fully connected layer 1 softmax
  - subject
  - relationship
  - object

- fully connected layer 2 softmax
  - \( \{a_{t,subj}\} \)
  - \( \{a_{t,rel}\} \)
  - \( \{a_{t,obj}\} \)

- fully connected layer 3 softmax

- three attention weights over each word
  - \( \{\tilde{a}_{t,subj}\} \)
  - \( \{\tilde{a}_{t,rel}\} \)
  - \( \{\tilde{a}_{t,obj}\} \)

- weighted average

- \( q_{subj} = \sum_{t=1}^{T} a_{t,subj} e_t \)

- \( q_{rel} = \sum_{t=1}^{T} a_{t,rel} e_t \)

- \( q_{obj} = \sum_{t=1}^{T} a_{t,obj} e_t \).
Localization module

Local spatial feature:

\[ x_s = \left[ \frac{x_{\text{min}}}{W_I}, \frac{y_{\text{min}}}{H_I}, \frac{x_{\text{max}}}{W_I}, \frac{y_{\text{max}}}{H_I}, \frac{S_b}{S_I} \right] \]

- \( x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}} \): coordinates of bounding box
- \( W_I, H_I \): width and height of image
- \( S_I, S_b \): area of image and bounding box
Localization module

\[ \tilde{x}_{v,s} = W_{v,s}x_{v,s} + b_{v,s} \]

\[ z_{loc} = \tilde{x}_{v,s} \odot q_{loc} \]

\[ \hat{z}_{loc} = z_{loc}/\|z_{loc}\|_2 \]

\[ s_{loc} = w_{loc}^T\hat{z}_{loc} + b_{loc} \]
Local spatial feature:

\[ x_s = \begin{bmatrix} x_{\text{min}} & y_{\text{min}} & x_{\text{max}} & y_{\text{max}} & S_b \end{bmatrix} \]

- \( x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}} \): coordinates of bounding box
- \( W_i, H_i \): width and height of image
- \( S_i, S_b \): area of image and bounding box
Relationship module

\[
\begin{align*}
\tilde{x}_{s1,s2} &= W_{s1,s2}x_{s1,s2} + b_{s1,s2} \\
\hat{z}_{rel} &= \tilde{x}_{s1,s2} \odot q_{rel} \\
\hat{s}_{rel} &= \frac{\hat{z}_{rel}}{\|\hat{z}_{rel}\|_2} \\
s_{rel} &= w_{rel}^T \hat{s}_{rel} + b_{rel}.
\end{align*}
\]
Loss function

Ground truth subject and object

\[ \text{Loss}_{\text{strong}} = -\log \left( \frac{\exp(s_{\text{pair}}(b_{\text{subj-gt}}, b_{\text{obj-gt}}))}{\sum_{(b_i, b_j) \in B \times B} \exp(s_{\text{pair}}(b_i, b_j))} \right) \]

Ground truth subject only

\[ \text{Loss}_{\text{weak}} = -\log \left( \frac{\exp(s_{\text{subj}}(b_{\text{subj-gt}}))}{\sum_{b_i \in B} \exp(s_{\text{subj}}(b_i))} \right) \]
Synthetic dataset

Shapes in different colors and size 5 by 5 grid

expression="the green square right of a red circle"

[subject] [relationship] [object]
Experiments on synthetic dataset

VGG-16 pretrained on ImageNet classification
Bounding box proposals are 5 by 5 grids
Experiments on synthetic dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline (loc module)</td>
<td>46.27%</td>
</tr>
<tr>
<td>our full model</td>
<td>99.99%</td>
</tr>
</tbody>
</table>

Baseline model: localization without relationship
Full model: localization and relationship
Qualitative results on synthetic dataset

expression = “the green square right of a red circle”

(baseline) $s_{loc}$

$s_{subj}$

$s_{obj}$
Visual Genome dataset

Real images and relationship annotations
Multiple expressions and bounding boxes on the images
Experiments on Visual Genome dataset

Fc7 output of a Faster-RCNN
VGG-16 pretrained on MSCOCO detection dataset
Experiments on Visual Genome dataset

Baseline model: localization without relationship
Full model: localization and relationship
Subject only: find the correct subject
Subject-object: find the correct subject-object pair

<table>
<thead>
<tr>
<th>Method</th>
<th>training supervision</th>
<th>P@1-subj</th>
<th>P@1-pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>subject-GT</td>
<td>41.20%</td>
<td>-</td>
</tr>
<tr>
<td>baseline</td>
<td>subject-object-GT</td>
<td>-</td>
<td>23.37%</td>
</tr>
<tr>
<td>our full model</td>
<td>subject-GT</td>
<td>43.81%</td>
<td>26.56%</td>
</tr>
<tr>
<td>our full model</td>
<td>subject-object-GT</td>
<td>44.24%</td>
<td>28.52%</td>
</tr>
</tbody>
</table>
Qualitative results on Visual Genome dataset

<table>
<thead>
<tr>
<th>ground-truth</th>
<th>our prediction</th>
<th>attention weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>expression=&quot;tennis player wears shorts&quot;</td>
<td>![Ground-truth Image]</td>
<td>![Attention Weights Chart]</td>
</tr>
</tbody>
</table>

Performance of language representation
Qualitative results on Visual Genome dataset

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “tennis player wears shorts”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image1" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image2" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image3" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “building behind bus”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image4" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image5" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image6" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “car has tail light”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image7" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image8" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image9" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “window on front of building”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image10" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image11" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image12" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “business name on sign”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image13" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image14" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image15" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “board on top of store”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image16" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image17" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image18" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “wine bottle next to glasses”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image19" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image20" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image21" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “marker on top of ledge”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image22" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image23" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image24" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
<th>Attention Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression</strong>: “chair next to table”</td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image25" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image26" alt="Image" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image27" alt="Image" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Google-Ref dataset

Real images and relationship expression
Only have ground truth bounding box for the subject

expression="a bear lying to the right of another bear"
Experiments on Google-Ref dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mao et al. [20]</td>
<td>60.7%</td>
</tr>
<tr>
<td>Yu et al. [30]</td>
<td>64.0%</td>
</tr>
<tr>
<td>Nagaraja et al. [21]</td>
<td>68.4%</td>
</tr>
<tr>
<td>baseline (loc module)</td>
<td>66.5%</td>
</tr>
<tr>
<td>our model (w/ external parser)</td>
<td>53.5%</td>
</tr>
<tr>
<td>our full model</td>
<td>69.3%</td>
</tr>
</tbody>
</table>

Same configuration on CNN and baseline model

Model (w/ external parser): using external language parser instead of language representation module
Qualitative results on Google-Ref dataset

<table>
<thead>
<tr>
<th>ground-truth</th>
<th>our prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>expression=&quot;a bear lying to the right of another bear&quot;</td>
<td>correct</td>
</tr>
<tr>
<td><img src="image1" alt="Ground-truth Image" /> <img src="image2" alt="Our Prediction Image" /></td>
<td><img src="image3" alt="Ground-truth Image" /> <img src="image4" alt="Our Prediction Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ground-truth</th>
<th>our prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>expression=&quot;man in sunglasses walking towards two talking men&quot;</td>
<td>correct</td>
</tr>
<tr>
<td><img src="image5" alt="Ground-truth Image" /> <img src="image6" alt="Our Prediction Image" /></td>
<td><img src="image7" alt="Ground-truth Image" /> <img src="image8" alt="Our Prediction Image" /></td>
</tr>
</tbody>
</table>
## Qualitative results on Google-Ref dataset

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Our Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;a bear lying to the right of another bear&quot;</td>
<td>correct</td>
</tr>
<tr>
<td>&quot;man in sunglasses walking towards two talking men&quot;</td>
<td>correct</td>
</tr>
<tr>
<td>&quot;a picnic table that has a bottle of water sitting on it&quot;</td>
<td>correct</td>
</tr>
<tr>
<td>&quot;woman in a cream colored wedding dress cutting cake&quot;</td>
<td>correct</td>
</tr>
<tr>
<td>&quot;a man going before a lady carrying a cellphone&quot;</td>
<td>correct</td>
</tr>
<tr>
<td>&quot;pizza slice not eaten&quot;</td>
<td>incorrect</td>
</tr>
<tr>
<td>&quot;a full grown brown bear near a young bear&quot;</td>
<td>correct</td>
</tr>
<tr>
<td>&quot;black dog standing on all four legs&quot;</td>
<td>incorrect</td>
</tr>
<tr>
<td>&quot;chair being sat in by a man&quot;</td>
<td>correct</td>
</tr>
</tbody>
</table>
Q: What endangered animal is featured on the truck?
A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?
A: Onto 24 ¾ Rd.
A: Onto 25 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?
A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: Who is under the umbrella?
A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.

Q: Why was the hand of the woman over the left shoulder of the man?
A: They were together and engaging in affection.
A: The woman was trying to get the man’s attention.
A: The woman was trying to scare the man.
A: The woman was holding on to the man for balance.

Q: How many magnets are on the bottom of the fridge?
A: 5.
A: 2.
A: 3.
A: 4.
Multiple choices questions start with which.

question="Which wine glass is in the man’s hand?"
Experiments on Visual-7W dataset

Fc7 output of a Faster-RCNN VGG-16 pretrained on MSCOCO detection dataset
RPN in Faster-RCNN to proposal regions for object
Experiments on Visual-7W dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhu et al. [32]</td>
<td>56.10%</td>
</tr>
<tr>
<td>baseline (loc module)</td>
<td>71.61%</td>
</tr>
<tr>
<td>our model (w/ external parser)</td>
<td>61.66%</td>
</tr>
<tr>
<td>our full model</td>
<td>72.53%</td>
</tr>
</tbody>
</table>

Same configuration on CNN and baseline model
Model (w/ external parser): using external language parser instead of language representation module
Qualitative results on Visual-7W dataset

question="Which wine glass is in the man’s hand?"

correct

question="Which person is wearing a helmet?"

correct
Qualitative results on Visual-7W dataset

question="Which wine glass is in the man’s hand?"
correct

question="Which person is wearing a helmet?"
correct

question="Which mouse is on a pad by computer?"
correct

question="Which head is that of an adult giraffe?"
correct

question="Which pants belong to the man closest to the train?"
correct

question="Which white pillow is leftmost on the bed?"
correct

question="Which red shape is on a large white sign?"
correct

question="Which is not a pair of a living canine?"
incorrect

question="Which hand can be seen from under the umbrella?"
correct
Comments

Contribution:
Compositional module for subjects, objects, relations
Attention mechanism to split the expression

Limitation:
Need special format of the referential expression
Potential Extensions

Potential Changes:
  Changing relationships to sentences
  (using a dataset with frames and learning just a FC layer for each role)
  (Or learning the frames)
Potential Improvements:
  Better visual feature extractors
Thanks
THANKS!